

DOCUMENT RESUME

ED 261 084

TM 850 499

AUTHOR Cason, Gerald J.; And Others
TITLE Handicapping Raters for Fairer Clinical Grading: A Practical Application.
PUB DATE Apr 85
NOTE 20p.; Paper presented at the Annual Meeting of the Association of Medical Colleges, Southern Group on Medical Education (San Juan, Puerto Rico, April 17-19, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests; Behavior Rating Scales; *Clinical Experience; *Grading; Higher Education; *Interrater Reliability; Medical School Faculty; *Medical Students; *Practicum Supervision; Student Evaluation; *Weighted Scores

ABSTRACT

To minimize the effects of systematic differences in raters' standards of clinical competence, a handicapping system was applied to the ratings made by fourteen preceptors of 128 junior year medical students in a 6-week psychiatry clerkship at the University of Arkansas for Medical Sciences. The handicap of a preceptor was the difference between the grand mean of all ratings and the mean of that preceptors ratings. The handicap was added to the preceptor's raw ratings. Final clerkship grades were based on weighted scores on the National Board of Medical Examiners psychiatry test, local psychiatry tests, and the rating. The observed range of preceptor handicaps (9 points) was quite large when compared to the observed range of raw ratings (16 points). Results showed that the adjusted ratings increased rater reliability by 25 to 35 percent. Adjusted ratings produced an increased correlation among the measures and a slight improvement in the adjusted total reliability when compared to the raw total. Thirteen percent of the students' grades changed one letter, depending on whether raw or adjusted scores were used.
(Author/GDC)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *

ED261084

HANDICAPPING RATERS FOR FAIRER CLINICAL GRADING:

A PRACTICAL APPLICATION¹

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☐ This document has been reproduced as
received from the person or organization
originating it.

☒ Minor changes have been made to improve
reproduction quality.

* Points of view or opinions stated in this docu-
ment do not necessarily represent official NIE
position or policy.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Cason, Gerald J.

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

JM 850 499

Handicapping Raters for Fairer Clinical Grading:
A Practical Application¹

Gerald J. Cason, Ph.D., Office of Educational Development
John L. Delk, Ph.D., College of Medicine
William G. Reese, M.D., College of Medicine
University of Arkansas for Medical Sciences

Abstract. Convincing evidence has accumulated which justifies adjusting clinical and ward ratings to remove the effects of systematic differences in rater standards and thereby achieve more nearly valid and reliable measures of clinical competence. Many researchers have advocated handicapping or otherwise adjusting preceptors' ratings of medical students' performance, but no published reports of applying the approach to actual grading could be located.

Handicapping was applied to the ratings made by 14 preceptors on 128 Junior year Medical students in a 6 week long Psychiatry Clerkship. Each student was rated by the preceptor to whom the student had been randomly assigned. A student could receive a raw ward rating of from 8 to 40. The handicap of a preceptor was the difference between the grand mean of all ratings and the mean of the preceptor's ratings. Adjusted ratings were found by adding a preceptor's handicap to each of that preceptor's raw ratings. Final clerkship grades were assigned on the basis of the student's Weighted Adjusted Total, i.e., the weighted sum of scores on NBME Psychiatry test, local Adult and Child Psychiatry tests, and adjusted ward rating.

The observed range of preceptor handicaps (9 points) was quite large when compared to the observed range of raw ratings (16 points). Across a reasonable range of assumed values for the reliability of raw ratings, the reliability of adjusted ratings was improved from 25 to 35 per cent over the reliability of unadjusted ratings. This was equivalent to .7 to .9 times the improvement expected from doubling the number of independent raters (from 1 to 2) for each student. Adjusting ratings produced an increased correlation among the measures and a slight improvement in the reliability of the adjusted total when compared to the raw total. Estimated reliabilities for the local tests (.52 and .60) and the totals (.43 and .48) were much lower than desirable (.70). When the gross distribution of grades was held constant, 13 percent of students' grades changed (8 rose and 8 declined) by one letter, depending on whether raw or adjusted ward ratings were used. Questions and complaints from students regarding rating standards and practices of different preceptors declined from fairly common (prior to handicapping) to almost nil. Faculty have voiced no interest in abandoning adjusted ratings.

Procedures for removing the effect of differences in raters' standards should be routinely applied to ward, clinic, and other performance ratings. Handicapping may often be done with only modest effort. But it may not be all that is needed; improved inventories and rater training, longer written tests may also be needed. In educational evaluation, no less than in law, the appearance of fairness and justice are equally as important as the substance of the matter. Both the appearance and actual justice of grades earned in clinical clerkships can be improved by adjusting ratings for differences in raters' standards.

¹Presented at the Association of American Medical Colleges, Southern Group on Medical Education Annual Meeting, April 17-19, 1985, San Juan, Puerto Rico. Reprints: GJ Cason, UAMS-OED-595, 4301 W Markham, Little Rock AR 72205 USA.

TM 850 499

Handicapping Raters for Fairer Clinical Grading: A Practical Application¹

Gerald J. Cason, Ph.D., Office of Educational Development
John L. Delk, Ph.D., College of Medicine
William G. Reese, M.D., College of Medicine
University of Arkansas for Medical Sciences

Marienfeld and Reid (1) recently discussed the problem of what they call the "easy grader" in clinical clerkships. They focused primarily on excessive leniency of some faculty members subjective assessments of medical student performance. They presented documentation, gathered over a six year period, of very lenient subjective assessments on the part of two faculty members in a department of medicine clerkship. They reported that any student assigned to either of these two easy grading faculty members had an excellent chance of receiving a grade of honors in the clerkship. As Marienfeld and Reid point out, clerkship coordinators receive many complaints from students regarding the disparity in rating patterns and apparent standards of different faculty members. Students are acutely aware that the difference in a single letter grade in one clerkship might change their overall class standing and, thereby, alter their prospect for admission to the increasingly more competitive, desirable internship and residency programs.

While Marienfeld and Reid only briefly mention it, their data clearly illustrates another problem which they label "grade inflation" in faculty members' subjective ratings of student performance. Consistent with the experience of many clerkship directors, Marienfeld and Reid's data indicate a general tendency among most faculty to be somewhat lenient with respect to the stipulative definitions in their grading systems. Consequently, the overall average of the faculty members' ratings is in the "B" range rather than the "C". On the other hand, since the grade of "C" (or its equivalent) is the stipulated average grade in many American medical colleges according to their officially declared policies, a faculty member may feel obligated to adhere closely to this formal norm in rating student performance. By so doing, however, this rater becomes by comparison to the actual behavior of most faculty a stringent, or low grader. In terms of the equitableness of ratings, from the point of view of a student who happens to be assigned to him or her, the low grader presents an even greater problem than does the easy grader, and is certain to generate more hostility and resentment. From a measurement view point, extreme graders, whether high or low, contribute to the same problem, i.e., large inter-rater variability. Large inter-rater variability is a very significant source of error of measurement which is reflected in lower coefficients of rating generalizability (2), reliability (3) and validity (3).

Marienfeld and Reid (1) conclude their discussion with two recommendations. Apparently, neither of these has yet been fully implemented in their own clerkship. First, they recommend providing detailed feedback to faculty in "an annual review of faculty grading patterns" in order to increase awareness of grading differences and alter the tendencies of easy graders. Second, they suggest handicapping "the ratings of easy graders based on the the difference between their (individual) mean yearly ratings and that of the whole department". Marienfeld and Reid acknowledge that handicapping might meet with faculty resistance. This may suggest why when their

¹Presented at the Association of American Medical Colleges, Southern Group on Medical Education Annual Meeting, April 17-19, 1985, San Juan, Puerto Rico. Reprints: GJ Cason, UAMS-OED-595, 4301 W Markham, Little Rock AR 72205 USA.

first recommendation (i.e., feedback to raters) was actually applied in a different medicine clerkship for four years (2) and was subsequently shown to have largely failed to correct differences in rater standards (4,5), the second kind of procedure (i.e., handicapping) was only suggested but has not yet been applied to actual grading in either of these or any other clerkship, in so far as can be determined from published reports (e.g., 1,3,4,5).

This report describes a junior year clerkship in a department of psychiatry in which both of the recommended procedures have been practically and fully applied, the methods used, and the results including overall student and faculty affective reaction.

Method

During academic year 1983-84, 97 male and 31 female (N=128) Junior medical students rotated through the six week long clerkship offered by the UAMS College of Medicine's Department of Psychiatry. Students were randomly assigned to one of the 14 attending and resident psychiatrists for supervised clinical experience in one of the 3 cooperating clinical facilities. In almost all cases the student was evaluated by the single preceptor from whom he or she received supervision. Each preceptor rated from 4 to 16 students (mean 10.29; median 8; and mode 8). The preceptor assigned a rating on a 5-point scale (5 was "excellent"; 1 was "unsatisfactory") for each of 8 criteria including: Relating with patients, instructors, and other clinical personnel; Selection and collection of data; Recording of data; Interpretation of data; Design of treatment plan; and, Participation in patient management. A student's raw ward rating score could range from 8 to 40. Ratings were recorded on machine scannable rating forms which had been computer preprocessed to have all student identifying information as well as the rating criteria preprinted on them. Only ratings had to be manually entered on the forms by the preceptors. Computer generated reports provided feedback to students. Rater means and handicaps were manually computed.

A handicap score for each rater (H_i) was computed as the difference between the grand mean ($\bar{X}_g$) across ratings of all students and the mean for an individual rater ($\bar{X}_i$), that is

$$H_i = \bar{X}_g - \bar{X}_i.$$

The original, raw rating scores were adjusted by adding a rater's handicap (H_i) to each of his or her ratings.

Feedback was provided to each rater comparing the rater's mean with the distribution of other raters' means. Similar information illustrating the adjustment procedure but omitting personal identification of specific raters was provided to students.

For the purposes of assigning a letter grade to be recorded in the student's official transcript, overall performance in the psychiatry clerkship was defined as the Adjusted Weighted Total (X_{awt}), i.e., a weighted combination of observed scores on three tests and the adjusted rating:

$$X_{awt} = \sum (a_i X_i) + a_{ar} X_{ar}$$

The nominal weights (a_i) for the three tests and the adjusted ratings (a_{ar}) were chosen by the faculty to reflect their a priori intuitive judgement of the relative

importance and validity of the NBME Psychiatry subtest (X_1), a locally written Child Psychiatry test (X_2), a locally written Adult Psychiatry test (X_3), and the adjusted ward rating (X_{ar}) while presumably taking into account differences in scale and variability of each measure. The nominal weight for each measure used to calculate the Adjusted Weighted Total is provided in Table 1.

Final letter grades were determined by the course co-ordinator's (Delk's) post hoc examination of the Adjusted Weighted Total score distribution and his selection of a minimum score for each grade. The set of minimum scores were subject to the restriction that College of Medicine policy required that the grade point average (GPA) of the letter grades assigned to all students in any Freshman, Sophomore, or Junior course or clerkship must fall in the range 1.5 through 2.5; given that A=4, B=3, C=2, D=1, and F=0.

For the purpose of examining the effect of using adjusted ward ratings in this clerkship, a Raw Weighted Total (X_{rwt}) was computed to contrast with the Adjusted Weighted Total (X_{awt}) for each student:

$$X_{rwt} = \sum (a_i X_i) + a_{rr} X_{rr}$$

The nominal weights (a_i) for the tests were the same as were used to compute the Adjusted Weighted Total. However, the nominal weight (a_{rr}) for the raw (i.e., unadjusted) ward rating (X_{rr}) had to be determined post hoc. The weight for the raw ward rating (a_{rr}) was selected such that the part-total correlation for the raw rating-raw total was equal to the observed correlation for the adjusted rating-adjusted total. This procedure assured that any difference observed in the two totals could not arise from a difference in the contribution of the rating component of the two totals.

To compare the impact of using raw versus adjusted totals on letter grades assigned, minimum raw total scores for each letter grade were chosen post hoc such that an equal frequency of A's, B's, C's etc. were produced as were observed for the adjusted total score distribution. This restriction assured that any change in grades assigned individual students would arise from a cause other than the overall distribution of grades.

Results

Overall results included the following. The observed range of preceptor handicaps (9 points) was quite large when compared to the range of raw ratings (16 points) received by individual students. Across a reasonable range of assumed values for the reliability of raw ratings, the reliability of adjusted ratings was improved from 25 to 35 per cent over the reliability of unadjusted ratings. This was equivalent to .7 to .9 times the improvement expected from doubling the number of independent raters (from 1 to 2) for each student. Adjusting ratings produced a small but observable increase in the common variance among the tests and ward ratings. This was reflected in increased correlations among the measures and a slight improvement in the reliability of the adjusted total when compared to the raw total. Estimated reliabilities for the local tests (.52 and .60) and the totals (.43 and .48) were much lower than desirable (.70). When the gross distribution of grades was held constant, 13 percent of students' letter grades changed (8 rose and 8 declined) by one letter, depending on whether raw or adjusted ward ratings were used.

Handicaps for preceptor ratings of students' ward performance ranged from -3.61 to 5.82 with a mean of 0.21. When applied to the rating of a student of average performance (see Table 1), this could move the student's reported rating from the bottom to the top of the observed distribution of unadjusted ratings or visa versa depending on whether the student was evaluated by the most or least stringent preceptor.

Insert Table 1 about here.

Table 1 provides descriptive statistics on the students' performance on the various measures: tests, ratings, and totals. The nominal weight for the adjusted rating is greater than that for the raw rating. The weight for the raw rating was found by applying trial weights until the correlation between the raw rating and raw total was the same (to four significant digits) as was the correlation between the adjusted rating and adjusted total.

If the nominal weight for the raw rating had been the same as was used for the adjusted rating, the correlation between raw rating and total would have been too high (i.e., .72). Under this condition the raw total would have been over-determined by the raw rating. From Table 1 it can be seen that the standard deviation of the two totals are not equal. Had the weight of the raw rating been chosen to equalize the variance in the totals (i.e., .63), the correlation between the raw rating and raw total would have been too low (i.e., .69) and the raw total would have been under-determined by the raw rating. Supplemental analyses not reported in detail here showed that across this entire range of weights for the raw rating, the pattern of results reported below remained invariant.

Insert Table 2 about here

Table 2 reports the correlations between tests, ratings, and totals. Only two of the correlations are of questionable difference from zero (i.e., $p < .08$). These are the correlations between raw ratings and the NBME and Child Psychiatry tests. This result is consistent with the presence of greater error (i.e., lower reliability or validity) in the raw ratings. The effect of adjusting the ratings was to slightly raise their correlations with the NBME and local Child Psychiatry tests. Under Maatsch's (6) assumption of a single factor underlying both tests and ratings of clinical competence, the pattern of correlations in Table 2 indicates that adjusting ratings produced a more nearly factorially valid measure of performance. Supplementary factor analyses were performed separately on the two sets of measures from which totals were calculated to quantify this qualitative impression. In each analysis only one meaningful factor (i.e., having an eigenvalue equal to or greater than one) could be extracted. Both analyses resulted in the same communalities between the first factor and the NBME (.50), Child Psychiatry (.41), and Adult Psychiatry (.16) tests. The communality of the adjusted rating (.08) with principle factor of its set of measures was slightly larger than was the communality of the raw rating (.05) with the principle factor underlying the set of measures of which it was an element. The corresponding factor loadings of the adjusted and raw rating on the first factor were .29 and .23 respectively. The factor loadings are interpretable as the correlation between the factor and the original measure (i.e., the adjusted and raw ratings). The communalities are the squares of the factor loadings thus represent the amount of variance shared or accounted for. Even though only a single meaningful factor was extracted in each factor analysis, in neither case was there much direct support for Maatsch's hypothesis. Only about 45 percent of the trace i.e., variance in the original set of variables, was accounted for by the first

factor. However, this may only reflect the relatively low reliability of all the local measures.

The KR-20 reliability of the NBME Psychiatry subtest is reported in Table 1. The KR-20 internal consistency reliability is a special case of the more general Cronbach's alpha (7). Alpha computed for both the local tests is also given in Table 1. Both alpha's for the local tests are less than what is often recommended for teacher made achievement tests (i.e., .70). But as is often the case with hand scored tests, this was not known until long after any corrective action was practical. In future, the reliabilities of the local tests can be raised to the desired range through the simple expedient of writing longer tests as specified by the Spearman-Brown formula. Alpha was not computed on the ratings. It is inappropriate (even if often used) for ratings because it is spuriously raised by correlated rater errors from such factors as the halo effect, response tendencies, etc.

The most appropriate measure of rating and rater reliability is the intra-class correlation. Direct computation of this value requires at least two independent ratings per student. Only a single rating per student was available; therefore, we chose to estimate an approximate value for this reliability through plausible assumptions. Table 3 reports four intra-class correlation reliabilities (assuming each student's rating is from one rater not an average of several raters) for both raw and adjusted ratings. First, the upper bound of the reliability is inferred from the case in which all error in the rating is assumed to be corrected by the adjustment procedure. In this case, using Ebel's (8) formula "d" and the variances of the raw and adjusted ratings, the error mean square is 4.02; thus, the upper limit for reliability for the raw rating is .75. This upper bound is difficult to take seriously as more than an illustration and a point of departure for estimating a more plausible value. However, it does provide a plausible minimum value for the error mean square in Ebel's formula and an estimate of the reduction in error derived from adjusting ratings independent of other sources of error or their magnitude.

Insert Table 3 about here

Although Symonds (9) reported that single-rater reliability was typically about .6 for subjective assessments under typical circumstances, research on rating medical students' clinical performance, where little rater training was used, indicated reliabilities from .2 (10) to .6 (11) with most single rater reliabilities falling between .3 and .5 (2,3,12,13). Stillman (14) found that extensive rater training could raise the single-rater reliability to .85. Prior research also indicated that lower inter-rater reliabilities were obtained in the affective-interpersonal domain than in the cognitive-technical, problem solving domain (11,15). From the above it seemed reasonable to assume that the single rater reliabilities reported for a study of a Junior year Medicine clerkship (4) provided a conservative, plausible range of values (.08 to .40) for the reliability of raw ward ratings in the present study. The median (.24) reported in that study (4) was considered to be a very conservative single value to use as an approximate estimate of the reliability of raw ward ratings here. This median was near the bottom of values reported in general; therefore, it was low enough to take into account that there was no rater training and the inventory used in the present study was somewhat weighted toward the affective-interpersonal domain.

Assuming the reliabilities for raw rating were as given in Table 3 and the reduction in mean square was defined as explained above, Ebel's formula was used to compute reliabilities for corresponding adjusted ratings (based on a single

rating-rater per student). Across this whole range of assumed reliabilities for raw ratings, the reliability of the adjusted ratings improved over the reliabilities for the raw ratings by from 25 to 35 percent. Applying Spearman-Brown's formula (16), the improvement can be shown to be equivalent to from .7 to .9 times the effect of increasing the number of independent raters of each student from 1 to 2. The median assumed reliability for the raw rating and the corresponding reliability for adjusted ratings were included in Table 1 for purposes of concise summary of these results.

At the bottom right of Table 1, alpha reliabilities are reported for the two totals. These were computed by treating the totals as if they were tests defined as weighted sums of scores on items defined by the other measures. The reported values may be directly computed from weights and the item and total variances also reported in Table 1. Consistent with all the other results, the reliability of the adjusted total was slightly higher. Alpha is unquestionably the correct method for estimating the reliabilities for totals if there is a single factor underlying clinical competence. The low reliability of the total presumably traces, at least in part, to the relatively low reliabilities of the constituent measures. Even were the single factor hypothesis rejected, alpha is useful in determining how consistent the different measures were that contributed to the assignment of a uni-dimensional letter grade.

Insert Table 4 about here

Table 4 provides a concrete example of the effect of choosing raw or adjusted ratings to determine final letter grades in this specific Psychiatry clerkship. The symmetry in the table is an artifact of forcing the total distribution of letter grades to be the same. Of these 128 students, 8 made grades one letter higher and 8 made grades one letter lower using adjusted rather than raw ratings.

Reaction of faculty and students were only informally assessed. Questions and complaints from students to the course co-ordinator regarding rating standards and practices of different preceptors declined from fairly common (prior to introducing handicapping) to almost nil. Faculty have voiced no interest in abandoning adjusted ratings. However, no indication was found that providing information on the contrasts between different raters' standards caused any subsequent change in those standards. As reported in other studies (2), variations in rater standards appear fairly resistant to change by simply informing raters how their average ratings differ from those of other raters.

Discussion and Conclusions

The results reported here were not sufficient to justify either conceptually or empirically, the handicapping of preceptors' ratings of medical students' ward performance. That justification has been more than adequately provided in prior research and was the rationale for using handicapping in this clerkship. Those prior studies provided not only the justification for handicapping but the expectation that adjusted ratings would be in some sense better than unadjusted ones. The results obtained, while not definitive, were consistent with these expectations.

The estimates of the reliabilities for the local tests and assumed range of reliability for the ward ratings were lower than desired. However, given the probable bias both at the local (i.e., author) and journal (i.e., reviewer and editor) level against publishing studies reporting low reliabilities for measures in practical education settings, the values obtained here may be more representative of the majority of programs than is the median value in the published literature. It

was in part for this reason that a range of values were assumed for the reliability of the raw ratings in these analyses. The results supporting improved reliability in the adjusted ratings were found across the entire range. Handicapping raters improved the reliability of the resultant adjusted ratings to approximately the same extent as would have been obtained by including the rating of an additional independent rater for each student. This improvement while more modest than desired was achieved at very small cost.

It might seem easy to dismiss these results as trivial. Reliability of the total was improved only from .43 to .48. Only 13 percent of the students' clerkship grade moved up or down by one letter grade depending on the use of handicapping. But, the improvement in reliability is based on very conservative assumptions. It might be larger than estimated here. Even if it were not, the upper limit on validity is set by the measure's reliability. Given no external criterion on which to base a choice, the more reliable measure or procedure should always be used, no matter how slightly more reliable. Also, as Cason and Cason (3) have argued, there is good reason to believe handicapping preceptors' ratings contributes directly to improved convergent, construct validity of the ratings. Seen in the light that the adjusted ratings were both more valid and reliable (even if only slightly so), the fact that few student grades were changed was immaterial. If only one student's grade were changed, handicapping would remain the more valid procedure.

Individual faculty tend to confuse the measurement and evaluation aspects of subjective assessment. At the level of a single observer, judge, or rater the two are practically inseparable, although conceptually distinct. The possibility of a difference between determining how much of some competence is present (i.e., measurement) and the merit or acceptability of different levels of competence (i.e., evaluation) does not usually arise when both are determined by a single person. Indeed, when only a single rater is responsible for the assessment, he or she ordinarily uses a single set of standards to serve both purposes.

When several faculty are responsible for assessing student performance the apparently arcane distinction between measurement and evaluation becomes very important. If each student is assessed by only a subset of the faculty, the distinction may become critical to the reliability and validity of the assessment. Each faculty member is responsible for reporting the quality of the students' performance which they have observed directly. To the extent that different faculty use different standards, their reports, as measures of student performance, are analogous to the readings on a set of uncalibrated thermometers. The process of adjusting the values provided by different instruments such that they conform to a standard scale concerns only the measurement process. It leaves undecided the evaluation issue, i.e., the standard to which all the instruments are to be calibrated or the standard of interpretation to be adopted.

At the evaluation level the faculty have joint responsibility to establish mutually acceptable, uniform standards for interpreting the measures of performance provided by ratings. The handicapping process permits each rater to use whatever standards he or she wishes in the role of measurement instrument while still assuring that the final interpretation will be based on a uniform standard for all students. Acceptance of handicapping by faculty requires at least tacit understanding of the distinction between their individual role in measuring student performance and their acceptance of the faculty's joint obligation to use a consistent set of evaluation standards across all students. Given this, handicapping can not be rationally opposed in principle. However a vigorous contest of values may remain between what Littlefield (5) has called the hawks and doves to decide what set of uniform

standards will be applied after ratings have been calibrated by handicapping or other methods. Ideally, the standards chosen will conform to those stipulated in formal policy statements. In our college, this decision was constrained by a college wide policy stipulating the possible range of the average of the grades assigned in any course. Final selection of the cutting scores for letter grades was delegated to the course co-ordinator, subject to the concurrence of the departmental chair and course faculty.

Handicapping or other procedures for removing the effect of differences in raters' standards should be routinely applied to ward, clinic, and other performance ratings of medical students, residents, and physicians. But, as was the case here, such handicapping should not be assumed to be all that is needed. Improved performance inventories and rater training may also be needed to bring inter-rater agreement into the desired range. Longer and or better developed written tests may be needed to assure that they make an equally useful contribution to the evaluation process.

In educational evaluation, no less than in law, the appearance of fairness and justice are equally as important as the substance of the matter. Both the appearance and actual justice of grades earned in clinical clerkships can be improved by adjusting ratings for differences in raters' standards. Furthermore, while other procedures may also be needed, the resources for other approaches to make improvements (such as time to train raters, write longer tests, etc) may not be easily available. By contrast, handicapping can be done with only modest extra effort on the part of the course director or faculty. In addition, handicapping is useful in any case where rater standards vary and students are rated by different raters, even if unadjusted ratings have unusually high but less than perfect inter-rater agreement ($.80 < r < 1.0$).

References

1. Marienfeld RD, & Reid JC. (1984). Six-year documentation of the easy grader in the medical clerkship setting. J Med Ed (59), 589-591.
2. Littlefield JH, Harrington JT, Anthracite NE, & Garman RE. (1981). A description and four-year analysis of a clinical clerkship evaluation system. J Med Ed (56), 334-340.
3. Cason GJ, & Cason CL. (1984). A deterministic theory of clinical performance rating: Promising early results. Eval and the Health Professions, 7(2), 221-247.
4. Cason GJ, Cason CL, & Littlefield JH. (1983). Controlling rater stringency error in clinical performance rating: Further validation of a performance rating theory. ERIC ED-228-314.
5. Littlefield JH, Ellis R, Cohen P, & Herbert R. (1984). Leniency and score distribution differences among clinical raters. RIME Proceedings, 199-203.
6. Maatsch JL, Huang RR, Downing S, Barker D, & Munger B. (1984). The predictive validity of test formats and a psychometric theory of clinical competence. RIME Proceedings, 76-80.
7. Cronbach LJ. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 297-334.
8. Ebel RE. (1951). Estimation of reliability of ratings. Psychometrika, 16, 407-424.
9. Symonds PM. (1931). Diagnosing Personality and conduct. New York: Century.
10. Pierlioni RG, Clark GM, & Dudding BA. (1979). A comparison of faculty, resident and nurse practitioner ratings fo ambulatory pediatric students. AERA.
11. Dielman TE, Hull AL, & Davis WK. (1980). Psychometric properties of clinical performance rating. Eval and the Health Professions, 3(1), 103-117.
12. O'Donohue WJ, & Wergin JF. (1980). Evaluation of medical students during a clinical clerkship in internal medicine. J Med Ed, (53), 55-58.
13. Printen KJ, Chappell W, & Whitney DR. (1973). Clinical performance evaluation of junior medical students. J Med Ed, (48), 343-348.
14. Stillman FL, Brown DR, Redfield DL, & Sabers DL. (1977). Construct validation of the Arizona Clinical Interview Rating Scale. Eval and Psychol Meas., 37, 1031-1038.
15. Cason CL, Cason GJ, & Littlefield JH. (1983). Variation of intra-rater stringency in cognitive-technical and affective-interpersonal clinical performance domains. AERA.
16. Remmers HH, Shock NW, & Kelly EL. (1927). An empirical study of the validity of the Spearman-Brown formula as applied to the Purdue Rating Scale. J Ed Psych, 18, 187-195.

Table 1. Junior Year Psychiatry Clerks Performance (N = 128);
Nominal Weights for Measures Contributing to Raw Total and Adjusted Total; and,
Estimates of Reliability of Separate Measures and Totals

Measure		Range	Mean	SD	Nominal Weight	Reliability ^a
NBME Psychiatry Subject Test	(X ₁)	255.00-755.00	531.60	85.65	.03	.75
Local Child Psychiatry Test	(X ₂)	49.00- 74.00	64.71	4.59	.23	.52
Local Adult Psychiatry Test	(X ₃)	69.00- 90.00	80.68	4.65	.22	.60 ^b
Raw (Unadjusted) Ward Rating	(X _{rr})	24.00- 40.00	32.90	4.33	.68	.24 ^b
Adjusted Ward Rating	(X _{ar})	23.15- 42.08	32.96	3.84	.71	.31 ^c
Raw Weighted Total	(X _{rwt})	57.37- 80.91	70.98	5.07		.43
Adjusted Weighted Total	(X _{awt})	60.46- 82.24	72.24	5.03		.48

^aReliabilities are alpha coefficients of internal consistency unless otherwise noted.

^bMedian intra-class correlation for single rater from earlier study (Cason, Cason, & Littlefield, 1983) used as best estimate.

^cIntra-class correlations for single rater based in part on assumed value of raw rating reliability.

Table 2. Inter-correlations Among Separate Measures, Raw (Unadjusted) and Adjusted Weighted Total Clerkship Performance (N=128)

		Child (X_2)	Adult (X_3)	Raw Rating (X_{rr})	Adjusted Rating (X_{ar})	Raw Total (X_{rwt})	Adjusted Total (X_{awt})
NBME Psychiatry Subject Test	(X_1)	.47	.28 ^a	.13 ^c	.17 ^b	.73	.76
Local Child Psychiatry Test	(X_2)		.23 ^a	.14 ^c	.17 ^b	.57	.59
Local Adult Psychiatry Test	(X_3)			.18 ^b	.18 ^b	.50	.50
Raw (Unadjusted) Ward Rating	(X_{rr})				.89	.71	.61
Adjusted Ward Rating	(X_{ar})					.67	.71
Raw Weighted Total	(X_{rwt})						.96

Note: Correlations different from zero at $p < .001$ unless otherwise indicated.

^a $p < .01$

^b $p < .05$

^c $p < .08$

Table 3. Approximations of Intra-class Correlation, Single Rater Reliabilities for Raw and Adjusted Ward Ratings in a Psychiatry Clerkship

	Raw Rating ^a	Adjusted Rating
Low	.08	.10
Median	.24	.31
High	.40	.54
Upper Bound	.75 ^b	1.00

^aLow, medium, and high reliability values for raw ratings approximated by values reported for medicine clerkship (Cason, Cason, & Littlefield, 1983).

^bAssuming all error corrected by adjustment procedure.

Table 4. Transitions in Overall Clerkship Grade Resulting from
Using the Adjusted or Raw Weighted Total Score
as the Summative Evaluation Criterion^a

		Grades Based on Adjusted Total					Total
		A	B	C	D	F	
Grades Based on Raw Total	A	10	3				13
	B	3	39	3			45
	C		3	61	2		66
	D			2	2		4
	F						0
Column Total		13	45	66	4		128

^aNote: Zero values omitted from table. Adjusted totals used to assign actual grades. Minimum Raw totals for each grade chosen to cause same overall, marginal distribution as for adjusted totals.

Grade point values of grades were: A=4, B=3, C=2, D=1, and F=0.